Using regression analysis to establish the relationship between number of rooms and house price: in Melbourne, Australia

Abdelrhman Dameen – z5427841 a.dameen@student.unsw.edu.au

Master of Data Science

The University of New South Wales (UNSW)

ZZSC9001 Foundations of Data Science - Assignment #2

23 October 2022

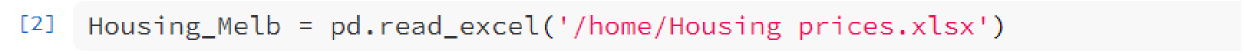
INTRODUCTION

We report here the results of a study examining the relationship between the number of rooms factor and the House price in Melbourne, Australia. A sample was randomly selected through the two-stage clustering method. [1] To illustrate how unique the data sample is, we will use multiple graphs and visual descriptions and show the results obtained using the multiple regression model. Then the model is used to predict a single house selling price in Melbourne east.
To examine the independence of variables, we carry out an F-test to determine whether there is any jointly significant relationship and how the variables affect each other.
To test and analyse the regression model again, we use the joint regression and then show the significance and implication of the P-value for the joint term.
The report concludes by selecting the most appropriate model for the dataset.
Python programming language is used in this report.

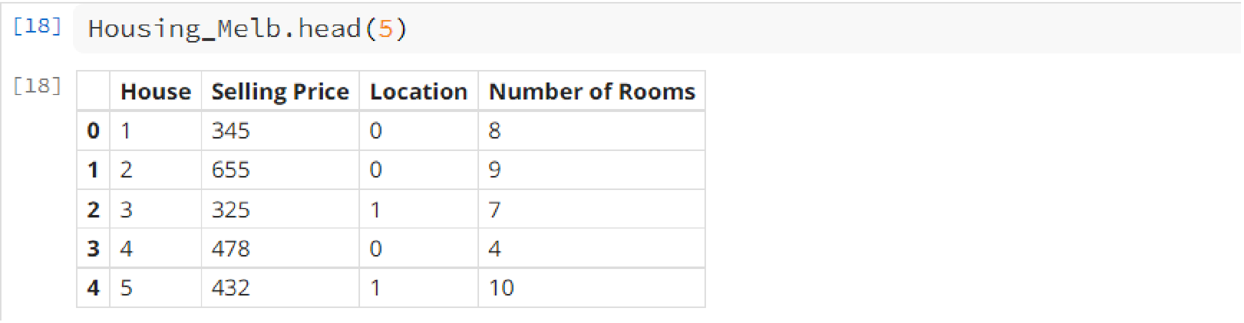
Method of Sampling and Assumptions

For a specific random sample, we use two-stage cluster sampling, a simple random sample of clusters is selected and then a simple random sample is selected from the units in each sampled cluster. One of the primary applications of cluster sampling is called area sampling, where the clusters are counties, townships, and cities… [2]
This method is used to get samples from different suburbs that are geographically dispersed within Melbourne.
The suburbs are grouped into two different clusters (West and East), and out of these two groups two suburbs were selected (Dandenong and Sunshine), then 10 random houses were selected from each of these suburbs.
We had to assume that the houses were selected randomly, and also that the samples were completely impartial.
We also assume that the samples were truly representative of the population.
Lastly, we assume that the location terms “East” and “West” are used appropriately.

General overview of the dataset the Location is a binary value (0) for West and (1) for East.



```
[2]  Housing_Melb = pd.read_excel('/home/Housing prices.xlsx')
```

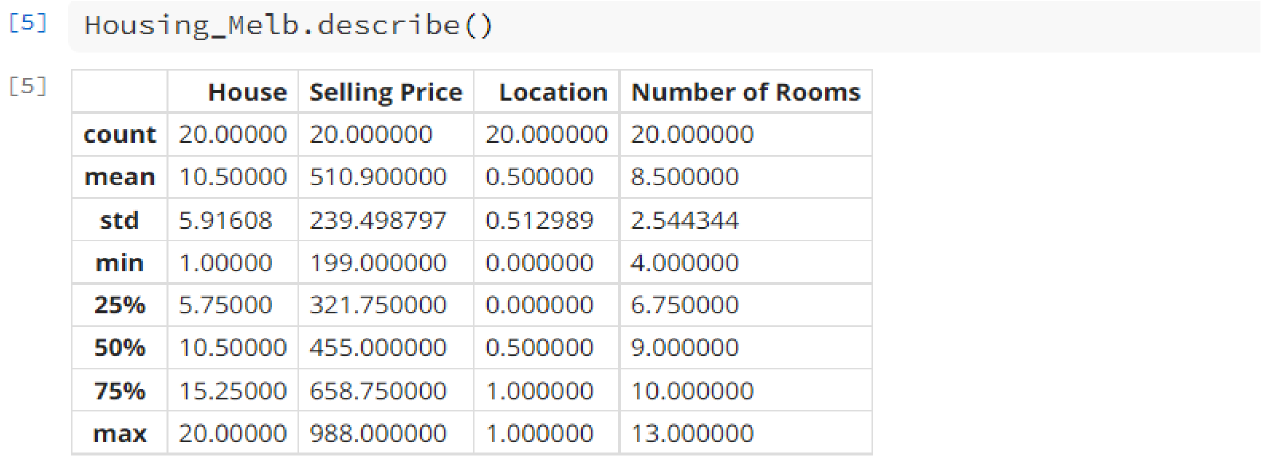


```
[18] Housing_Melb.head(5)
```

[18]

	House	Selling Price	Location	Number of Rooms
0	1	345	0	8
1	2	655	0	9
2	3	325	1	7
3	4	478	0	4
4	5	432	1	10

The general statistical description of the dataset.

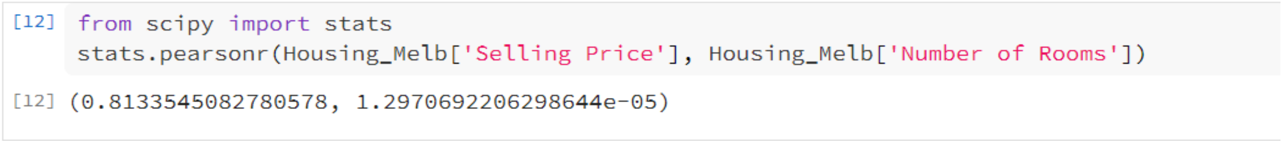


```
[5]  Housing_Melb.describe()
```

[5]

	House	Selling Price	Location	Number of Rooms
count	20.00000	20.000000	20.000000	20.000000
mean	10.50000	510.900000	0.500000	8.500000
std	5.91608	239.498797	0.512989	2.544344
min	1.00000	199.000000	0.000000	4.000000
25%	5.75000	321.750000	0.000000	6.750000
50%	10.50000	455.000000	0.500000	9.000000
75%	15.25000	658.750000	1.000000	10.000000
max	20.00000	988.000000	1.000000	13.000000

The relationship between the selling price and the number of rooms.



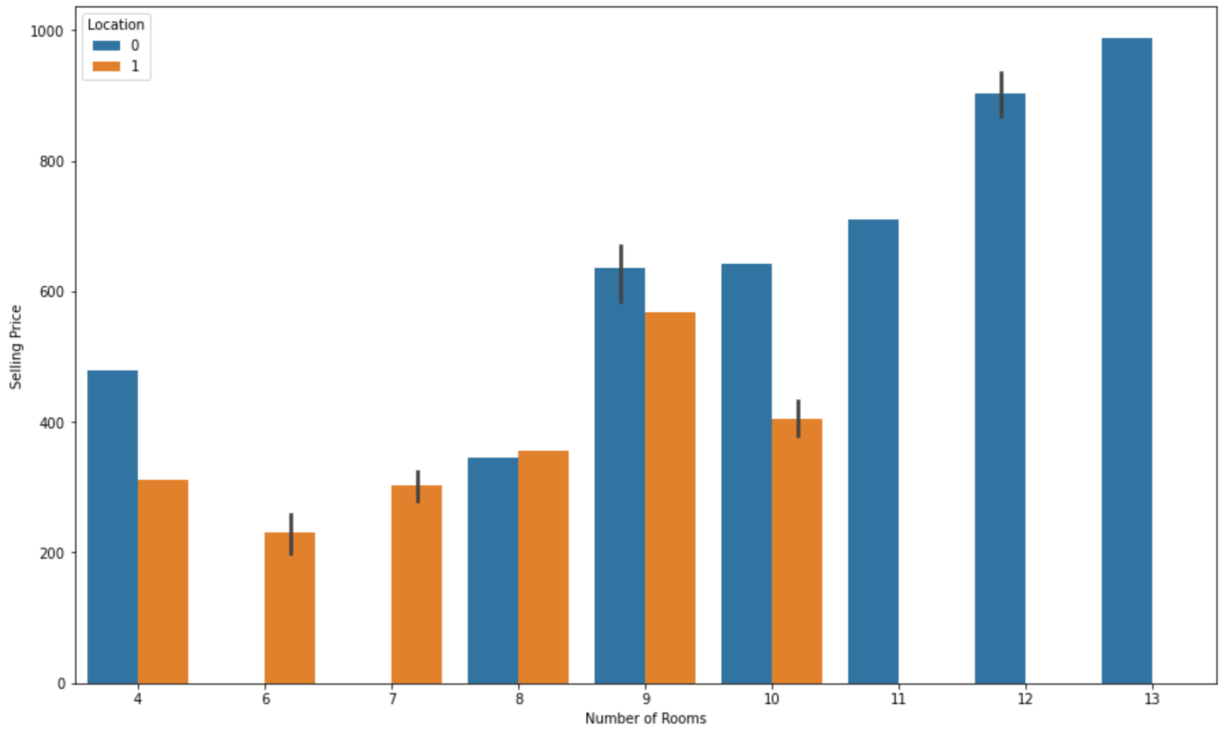
```
[12] from scipy import stats
     stats.pearsonr(Housing_Melb['Selling Price'], Housing_Melb['Number of Rooms'])

[12] (0.8133545082780578, 1.2970692206298644e-05)
```

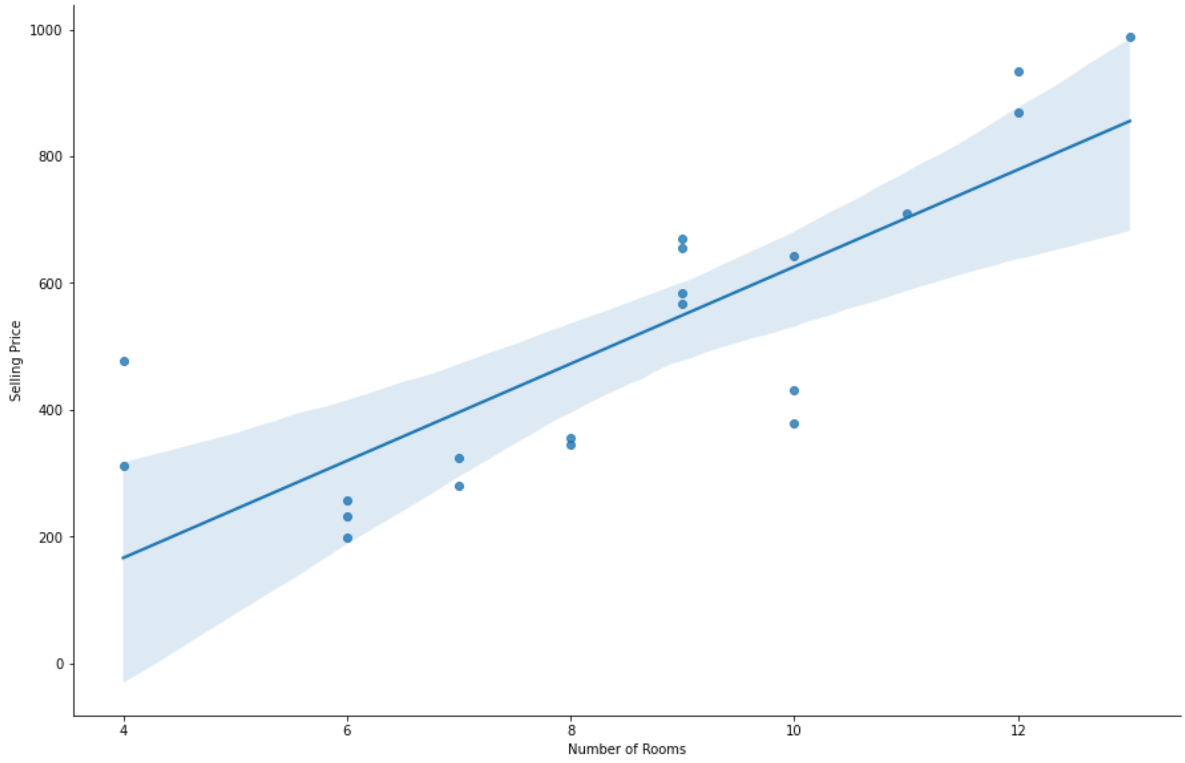
Regression works on the line equation, y = m * x + c, the trend line is set through the data points to predict the outcome. [3]

Graphs and sample data representation

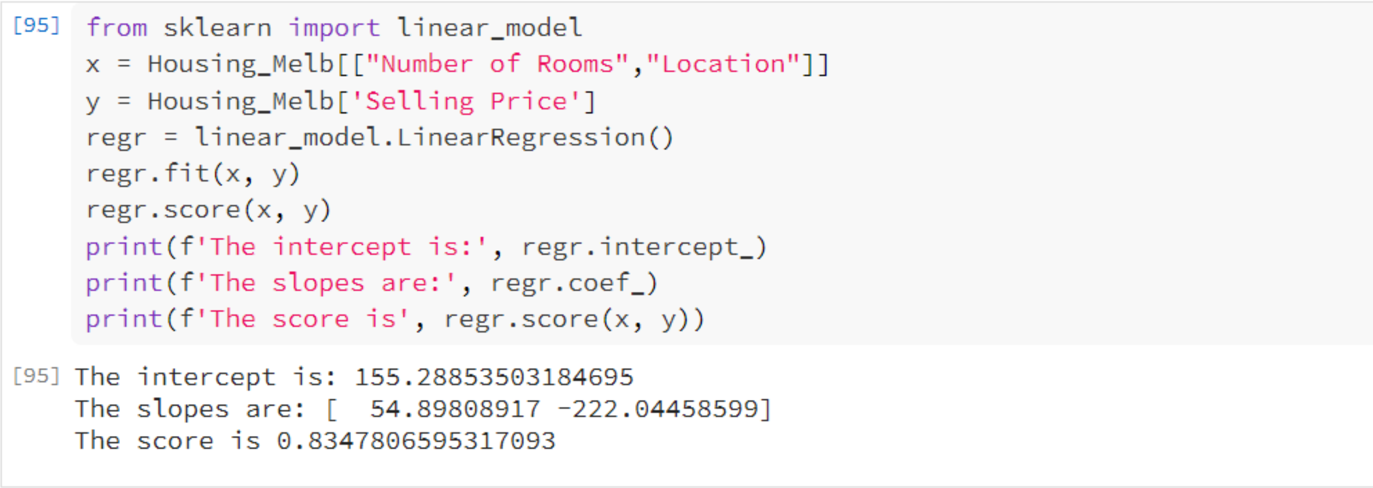
To represent the sample data in a graphical form, a bar plotting method has been used.
The blue bars represent the East side (Dandenong, marked 0) and the orange bars represent the West side (Sunshine, marked 1).
The reason for using this method is to give a clear view of the data from each side of the given location.



The correlation between the number of rooms and the selling price is (0.81335).
Clearly, the linear relationship between the two variables is very strong.
The plot below shows the linear relationship.
Looking at this visualization, one can easily and quickly find out what type of relationship we exists between the selling price and the size of the house.



Multiple regression equation and estimated selling price



```
[95] from sklearn import linear_model
     x = Housing_Melb[["Number of Rooms","Location"]]
     y = Housing_Melb['Selling Price']
     regr = linear_model.LinearRegression()
     regr.fit(x, y)
     regr.score(x, y)
     print(f'The intercept is:', regr.intercept_)
     print(f'The slopes are:', regr.coef_)
     print(f'The score is', regr.score(x, y))
```

```
[95] The intercept is: 155.28853503184695
     The slopes are: [  54.89808917 -222.04458599]
     The score is 0.8347806595317093
```

As per the code snippet above, we can see that intercept 155.29 and the two slopes are respectively 54.90 (number of rooms) and -222.04 (location).

Our equation will be $y = 155.29 + 54.9x_1 – 222x_2$

Estimated Selling Price = 155.29 + 54.9 (**Number of Rooms**) − 222.04 (**Location**)

Interpretation of slopes:

a) A slope of 54.9 for the number of rooms represents the estimated change in price (in thousands of dollars) for every increase of one room on either side (East and West)

b) A slope of -222.04 for location, represents the estimated change in price in the negative which means the West side (marked as 1) of the city has an average lower selling price of the house than houses on the East side (marked as 0) by 222.04.

House price prediction:

For a house with nine rooms located in Melbourne's East,

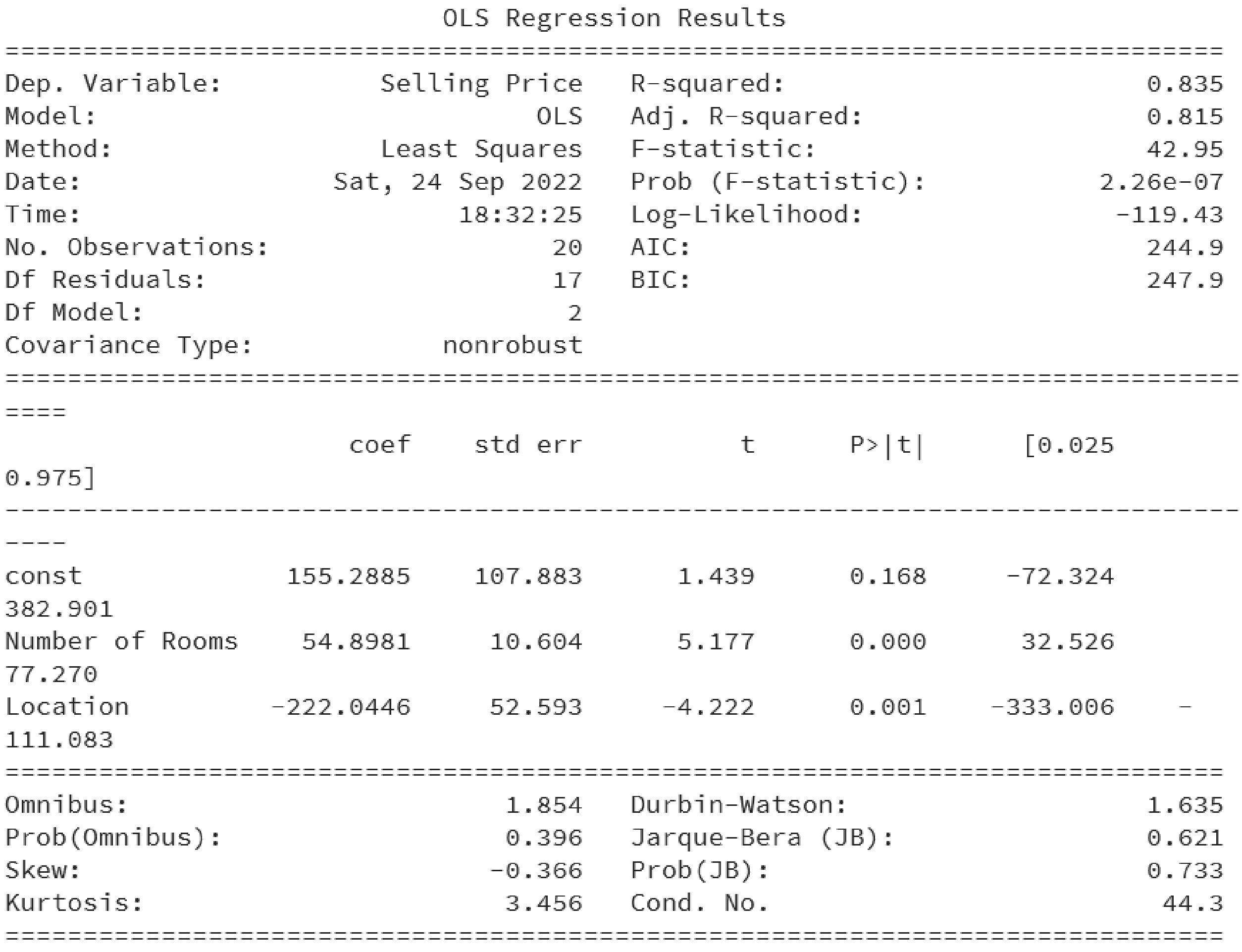
price prediction = 155.29+54.9(**9**) −222.04(**0**) = 649.39.

F statistic and the relationship between dependent and independent variables

The F-test of overall significance indicates whether our linear regression model provides a better fit to the dataset than a model that contains no independent variables.[4]

Ho: β1=β2=0 (No relationship between dependent and independent variables)

H1: = (At least one of the coefficients β is not zero)

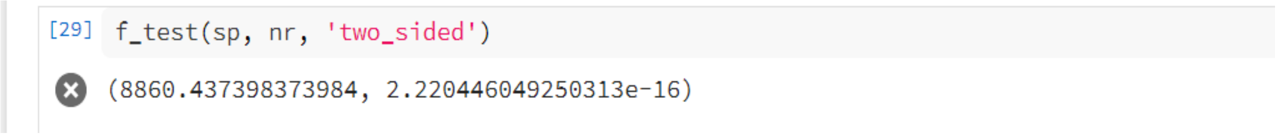


```
                            OLS Regression Results
==============================================================================
Dep. Variable:          Selling Price   R-squared:                       0.835
Model:                            OLS   Adj. R-squared:                  0.815
Method:                 Least Squares   F-statistic:                     42.95
Date:                Sat, 24 Sep 2022   Prob (F-statistic):           2.26e-07
Time:                        18:32:25   Log-Likelihood:                -119.43
No. Observations:                  20   AIC:                             244.9
Df Residuals:                      17   BIC:                             247.9
Df Model:                           2
Covariance Type:            nonrobust
===============================================================================
====
                      coef    std err          t      P>|t|      [0.025
0.975]
-------------------------------------------------------------------------------
----
const             155.2885    107.883      1.439      0.168     -72.324
382.901
Number of Rooms    54.8981     10.604      5.177      0.000      32.526
77.270
Location         -222.0446     52.593     -4.222      0.001    -333.006    -
111.083
==============================================================================
Omnibus:                        1.854   Durbin-Watson:                   1.635
Prob(Omnibus):                  0.396   Jarque-Bera (JB):                0.621
Skew:                          -0.366   Prob(JB):                        0.733
Kurtosis:                       3.456   Cond. No.                         44.3
==============================================================================
```

With an F-test of 42.95 and a P-value below <0.05 at a significant level below 5%, we reject the null hypothesis, because our coefficients have either positive or negative values which will impact our model. We can also see that there is a jointly significant relationship between the selling price and the other two independent variables based on our R-squared (83.48%), which means the 83% variation in the selling price could be explained by the changes in the independent variables.

The effect of the dependent variable on the independent variables

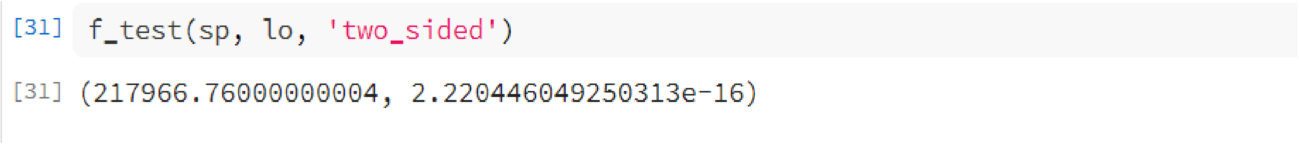
Checking the effect with selling price as the dependent variable, number of rooms as the independent variable



```
[29] f_test(sp, nr, 'two_sided')

     (8860.437398373984, 2.220446049250313e-16)
```

Checking the effect with selling price as the dependent variable, location as the independent variable



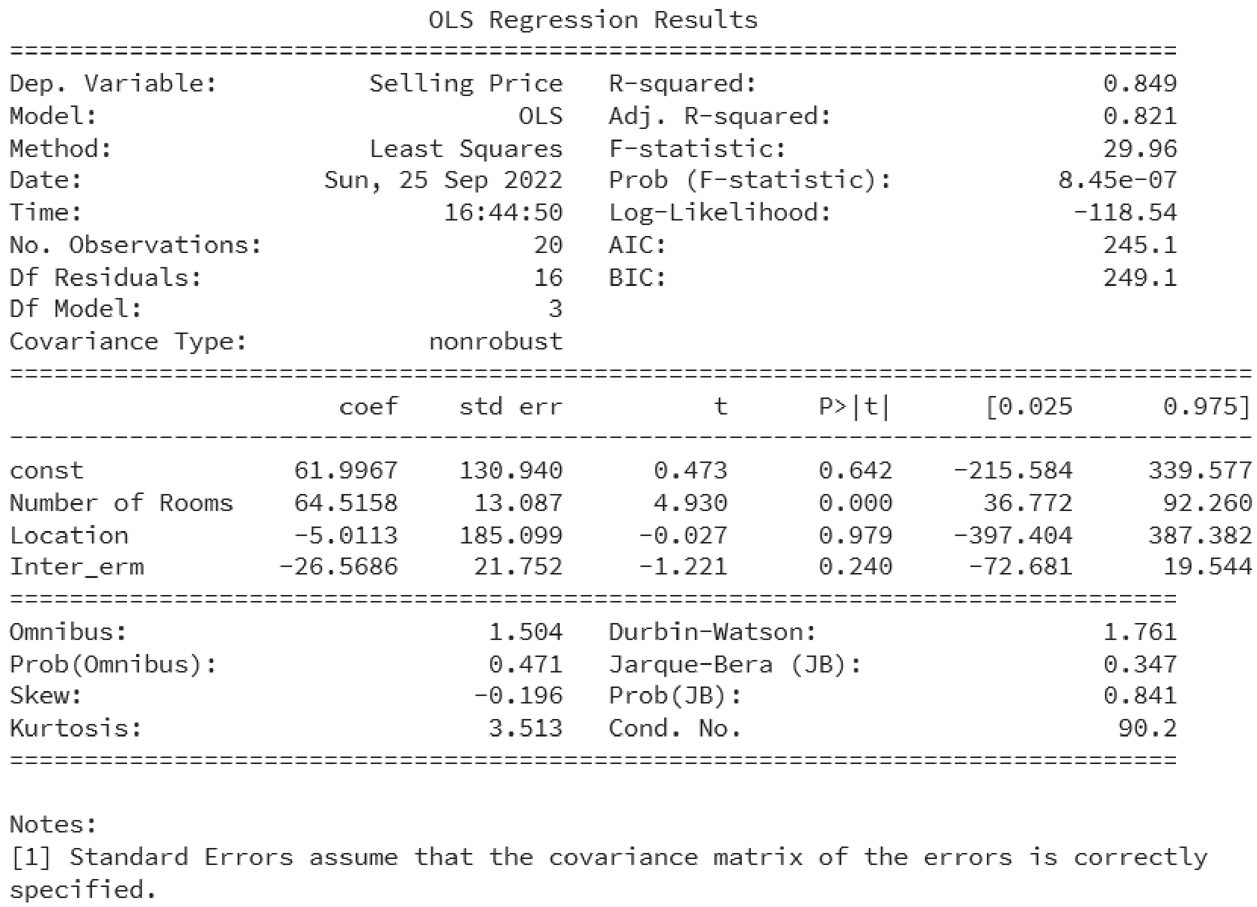
```
[31] f_test(sp, lo, 'two_sided')

[31] (217966.76000000004, 2.220446049250313e-16)
```

We can note that the P-value is less than 0.05 (at a 5% level of significance), so we can reject the null hypothesis.
This means a change in either the number of rooms or location affects the selling price.

Adding a joint term, X1X2 into the regression model and calculating the P-value.

The interaction term is calculated by multiplying the number of rooms by the location and running the new regression model.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:          Selling Price   R-squared:                       0.849
Model:                            OLS   Adj. R-squared:                  0.821
Method:                 Least Squares   F-statistic:                     29.96
Date:                Sun, 25 Sep 2022   Prob (F-statistic):           8.45e-07
Time:                        16:44:50   Log-Likelihood:                -118.54
No. Observations:                  20   AIC:                             245.1
Df Residuals:                      16   BIC:                             249.1
Df Model:                           3
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
const              61.9967    130.940      0.473      0.642    -215.584     339.577
Number of Rooms    64.5158     13.087      4.930      0.000      36.772      92.260
Location           -5.0113    185.099     -0.027      0.979    -397.404     387.382
Inter_erm         -26.5686     21.752     -1.221      0.240     -72.681      19.544
==============================================================================
Omnibus:                        1.504   Durbin-Watson:                   1.761
Prob(Omnibus):                  0.471   Jarque-Bera (JB):                0.347
Skew:                          -0.196   Prob(JB):                        0.841
Kurtosis:                       3.513   Cond. No.                         90.2
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

Estimated Selling Price = 61.997 + 64.516 (**Number of Rooms**) −5.011 (**Location**) −26.569 (**Number of Rooms * Location**)

This new model shows a higher P-value than the previous one (8.54), which implies that compared with the original model with a less negative coefficient and dependencies on the other variable (number of rooms), the location variable has a different effect on selling price.

Most Appropriate Model

Based on the analysis we can say that the regression model with a joint term is the most appropriate model because it has a higher number of adjusted R-squared values (0.8205) compared to the original (0.8153).
Here we get a better estimation of the selling price of houses on either side of the city with the joint term.

Recommendations / Conclusion

To improve the model and increase its efficiency, more variables should be added to the random sample.

Appendix

Appendix I: Access the main code here

Graph 3 [Multiple regression using the two different locations]
Graph 4 [Bar plot]
Graph 5 [Confusion matrix]
Graph 6 [Candlestick chart]
Graph 7 [Multiple Regression graph]

Appendix II: Reference

[1] Aldina Dervic & Linnea Ylinen in What determines housing prices? Bachelor thesis in economics See Here

[2] ‘Website’ QuestionPro Cluster Sampling: Definition, Method, and Examples See here

[3] The Regression Equation Introductory Statistics. Authored by: Barbara Illowski, Susan Dean. See here

[4] Statistics How To. Statistics for the rest of us! F-Statistics See here